IMPLEMENTACIÓN DE APACHE SPARK PARA ANÁLISIS DE PROTEÍNAS

HÉCTOR PATRICIO CASTILLO VÁSQUEZ
INGENIERO CIVIL EN COMPUTACIÓN

RESUMEN

Este proyecto se trata sobre el uso de tecnologías distribuidas para resolver problemas relacionados con la manipulación de datos masivos y complejos. El problema específico abordado es el de AFAL2, una aplicación web que permite analizar datos de proteínas, los cuales se obtienen del Protein Data Bank (PDB). AFAL2 permite consultar tres tipos de frecuencias. Además, este sistema funciona sobre una base de datos relacional, hecha en PostgreSQL y no tiene el mejor tiempo de respuesta frente a operaciones complejas. Además considerando que el Protein Data Bank crece constantemente, esto puede ser un problema a futuro. Para la investigación realizada fue necesario analizar la base de datos de AFAL2 y extraer las consultas que permiten responder a las frecuencias para luego replicar estas consultas usando una herramienta de programación distribuida, específicamente un framework llamado Apache Spark. Finalmente se realizaron pruebas para comparar la efectividad del algoritmo en Spark versus la base de datos hecha en PostgreSQL. Para el desarrollo de los algoritmos que responden a las frecuencias de AFAL2, se usó el modelo de programación Map Reduce, el cual consiste en dos operaciones, la primera es el map que separa los datos en los distintos nodos que componen el sistema distribuido para dividir el trabajo y más tarde unirlos en la etapa de Reduce. Replicar las consultas requirió el uso de varias operaciones Map y Reduce. Para probar la efectividad de los algoritmos con respecto a las consultas sql se realizaron varias pruebas en distintos ambientes, entre los que destacan notebook con PostgreSQL, notebook con Spark, y un cluster de Amazon EMR con Spark. En cuanto a los resultados, si bien no fueron los más favorables para Spark, se puede concluir que la tasa de crecimiento de tiempo de respuesta con respecto a un aumento en la cantidad de datos es mucho menor en Spark que en PostgreSQL, por lo que si en algún momento el Protein Data Bank crece lo suficiente, se podrán encontrar problemas lo suficientemente grandes como para aprovechar el potencial de Spark.